Assumptions in Dimensionality Reduction Techniques

(1) Principal Component Analysis (PCA) assumptions

Assumption 1 → Variables (Multiple variables) should be measured at the continuous level. Example of continuous variable are ratio or interval variables. [Continuous Data]

Assumption 2 → There needs to be Linear Relationship between all variables. The reason for this assumption is that PCA is based on Pearson Correlation Coefficient. [Pearson correlation Coefficient]

Assumption 3 → We should have sampling adequancy, which simply means PCA to produce reliable result, large enough sample size are required. Generally, a minimum of 150 cases or 5 to 10 cases per variable has been recommended as minimum sample size. [KMO measure of Sampling Adequancy]

Assumption 4 → Data should be suitable for data reduction. Effectively we need to have adequate correlations between variables in order for variables to be reduced to a smaller number of components. [Correlation]

Assumption 5 → No significant outliers. Outlier treatment are important because these can have a disproportionate influence on results. [Outlier treatment]

(2) Linear & Quadratic Discriminant Analysis (LDA, QDA) assumptions

Assumption 1 → Both LDA & QDA assume that the predictor variable X are drawn from multivariate Gaussian (aka normal) distribution.

Assumption 2 → LDA assumes equality of covariance among the predictors variable X across each all levels of Y. This assumption is relaxed with the QDA model.

Assumption 3 → LDA and QDA require the number of predictor variables (p) to be less than the sample size (n). A simple thumb rule to use LDA & QDA on datasets where $n \geq 5 \times p$

$n \geq 5$ multiply P

LDA is much more flexible classifier than QDA, so has substantially low variance. This potentially lead to improved prediction performance.

③ Factor Analysis Assumptions –

Assumption 1 → No outliers, assumes that there are no outliers in the data.

Assumption 2 → Adequate sample size, the case must be greater than factor.

Assumption 3 → No perfect multicollinearity, factor analysis is interdependency technique. There should not be perfect multicollinearity between variables.

Assumption 4 → Homoscedasticity, since factor analysis is a linear function of measured variable, it does not require homoscedasticity between variables.

Assumption 5 → Linearity, factor analysis is also based on linearity assumptions. Non-linear variable can also be used. After transfer, however it changes into linear variables.

Assumption 6 → Interval Data, Interval data are assumed.

④ T-SNE Assumptions –

Assumption 1 → Local structure of manifold is linear. The reason of this assumption is important is that the distance between neighboring points is measured in Euclidean distance, which assumes linearity.

Assumption 2 → t-SNE is non-deterministic. We can run it multiple times and get a different result each time.